All Tools Data Analysis Project

During our training at

National Telecommunication Insitute And ITIDA

Full Data Analysis Process with Python

Full Data Analysis Process with SQL

Full Data Analysis Process with Excel

Our Dataset Talk About

Project Goal: Analyze song performance, compare songs by genre, language,
duration, and explicit content, and extract valuable insights to support decision
making in the music industry

Key dimensions covered in the dataset include:

- **Song Attributes:** [song_title], [artist], [album], [genre], [language], [duration]
- **Performance Metrics:** [popularity], [stream]
- **Classifications:** [explicit_content], [popularity_level], [streams_level], [duration_minute], [date_group]
- **Production Data:** [composer], [producer]

Our Full Data Analysis Process with Excel

. Data cleaning

No duplication

- o Deal with outliers
  - ▪ There is outliers in [duration] and its count 349 it will be deleted
- o Deal with nulls
  - ▪ Filling nulls in [language] by mode
  - ▪ Filling nulls in [duration] by mean
  - ▪ Delete [collaboration] Because it contains many nulls = 35000

o Feature Engineering

- ▪ Create columns like
  - o popularity_level
  - o duration_minute
  - o streams_level
  - o date_group

▪ the columns that created will Facilitate analysis and comparison across different group

1-

Our Full Data Analysis Process with Excel

We also performed a complete data analysis process using Excel, leveraging pivot tables, slicers, and
dashboards to ensure data cleaning, interactive exploration, and clear visualization of insights from the Spotify_songs data set

2-**Pivot Tables & Slicers**
Pivot Tables Built For:

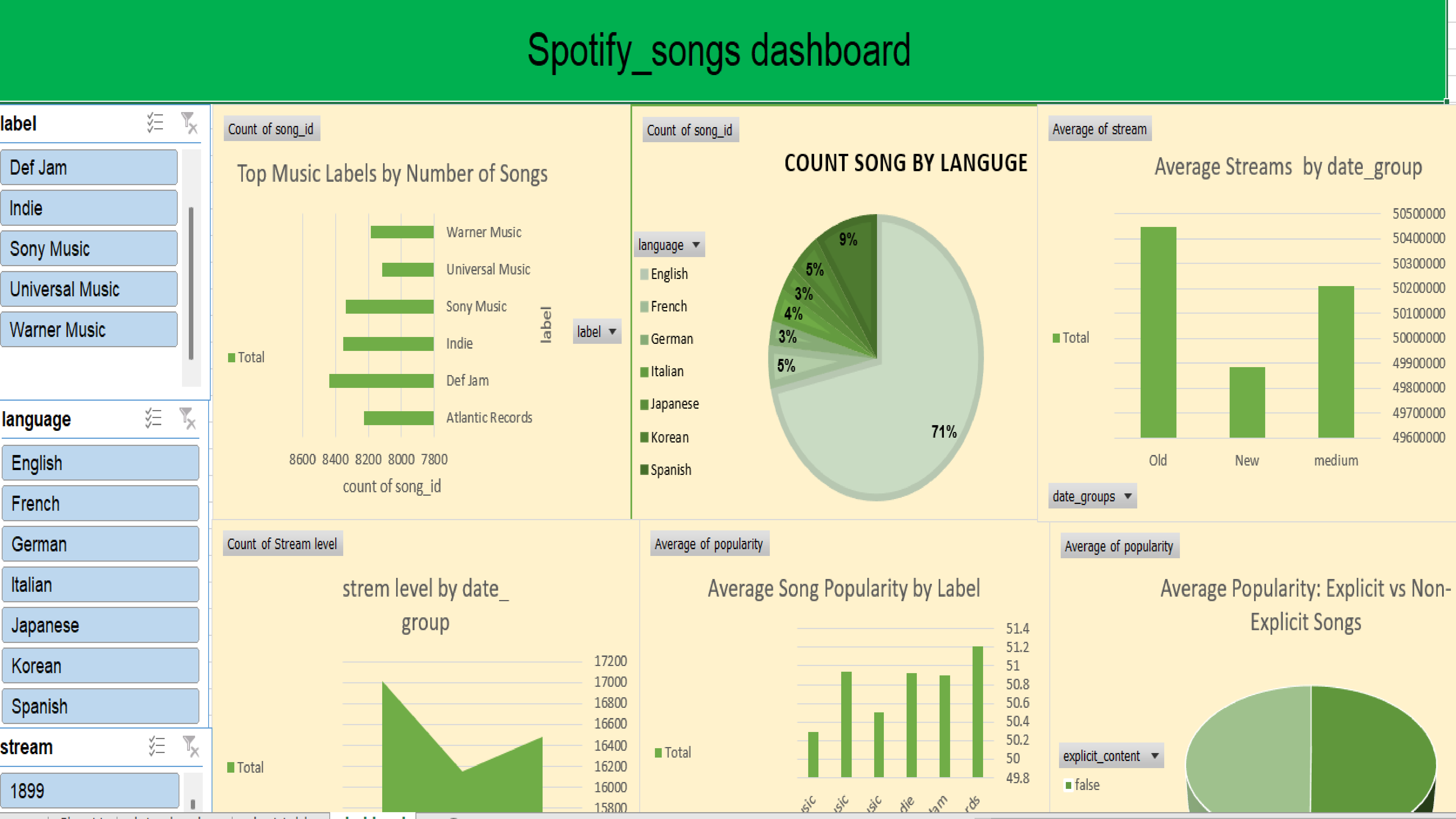
1) Top Music Labels by Number of Songs
2) Average Song Popularity by Label
3) Average Streams: Old vs Recent Songs
4) Average Popularity: Explicit vs Non-Explicit Songs

Slicers Added:

Our Full Data Analysis Process with Excel

•Key Insights

•**Sony Music** and **Universal Music** are the top labels in terms of number of songs.
•**English songs** dominate the dataset, representing about **71%** of all songs.
•**Older songs** achieved the highest number and average of **streams** compared to newer releases.
•**Recent songs** (new/medium) attract fewer streams than older ones.
•**Sony Music** and **Universal Music** have the highest average popularity, while **Indie labels** show lower popularity.
•There is **no significant difference** in popularity between **Explicit** and **Non-Explicit** songs.
•**English songs** are the most common language for **Explicit content**.
•The largest number of releases occurred in the **old period (17,016 songs)**, while recent years show fewer releases (**16,147 songs**).



Spotify_songs dashboard

Our Full Data Analysis Process with SQL

We also performed a complete data analysis process using
SQL **to ensure robust validation, efficient querying,**
and powerful insights from the Spotify_songs dataset.

DATA EXPLORATION

□ **Univariate Analysis:**

Studied each column separately. For numeric features (popularity, stream), we used **histograms** and **boxplots** to check distributions and outliers. For categorical features (genre, language, duration_minute, explicit_content, date_group, popularity_level, streams_level),

Our Full Data Analysis Process with SQL

Data cleaning

No duplication

- o Deal with outliers
  - ▪ There is outliers in [duration] and its count 349 it will be deleted
- o Deal with nulls
  - ▪ Filling nulls in [language] by mode
  - ▪ Filling nulls in [duration] by mean
  - ▪ Delete [collaboration] Because it contains many nulls = 35000

o Feature Engineering

- ▪ Create columns like
  - o popularity_level
  - o duration_minute
  - o streams_level
  - o date_group

▪ the columns that created will Facilitate analysis and comparison across different group

Our Full Data Analysis Process with SQL

Feature Engineering •

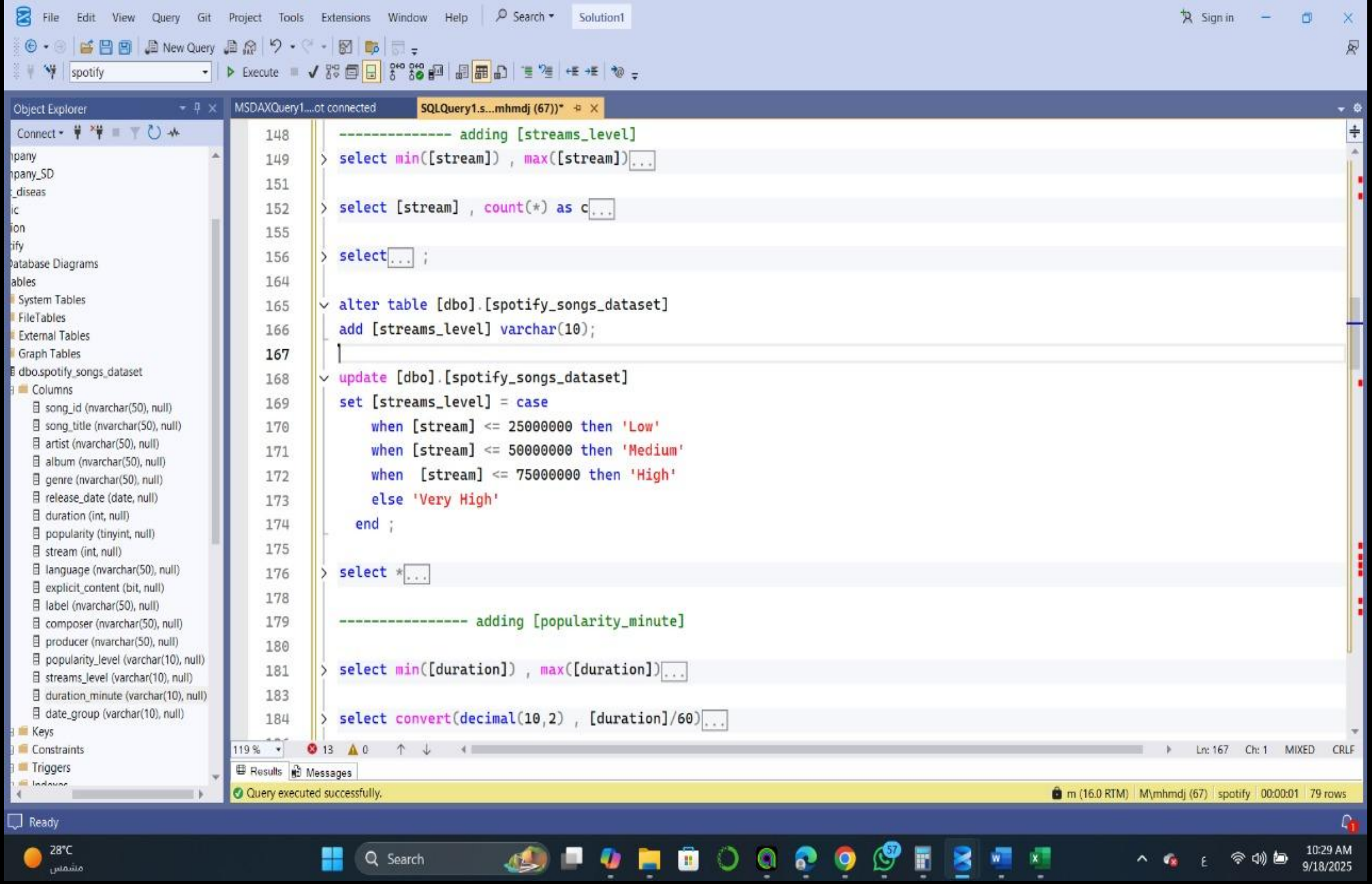
▪ Create columns like •

o popularity_level •

o duration_minute •

o streams_level •

o date_group •

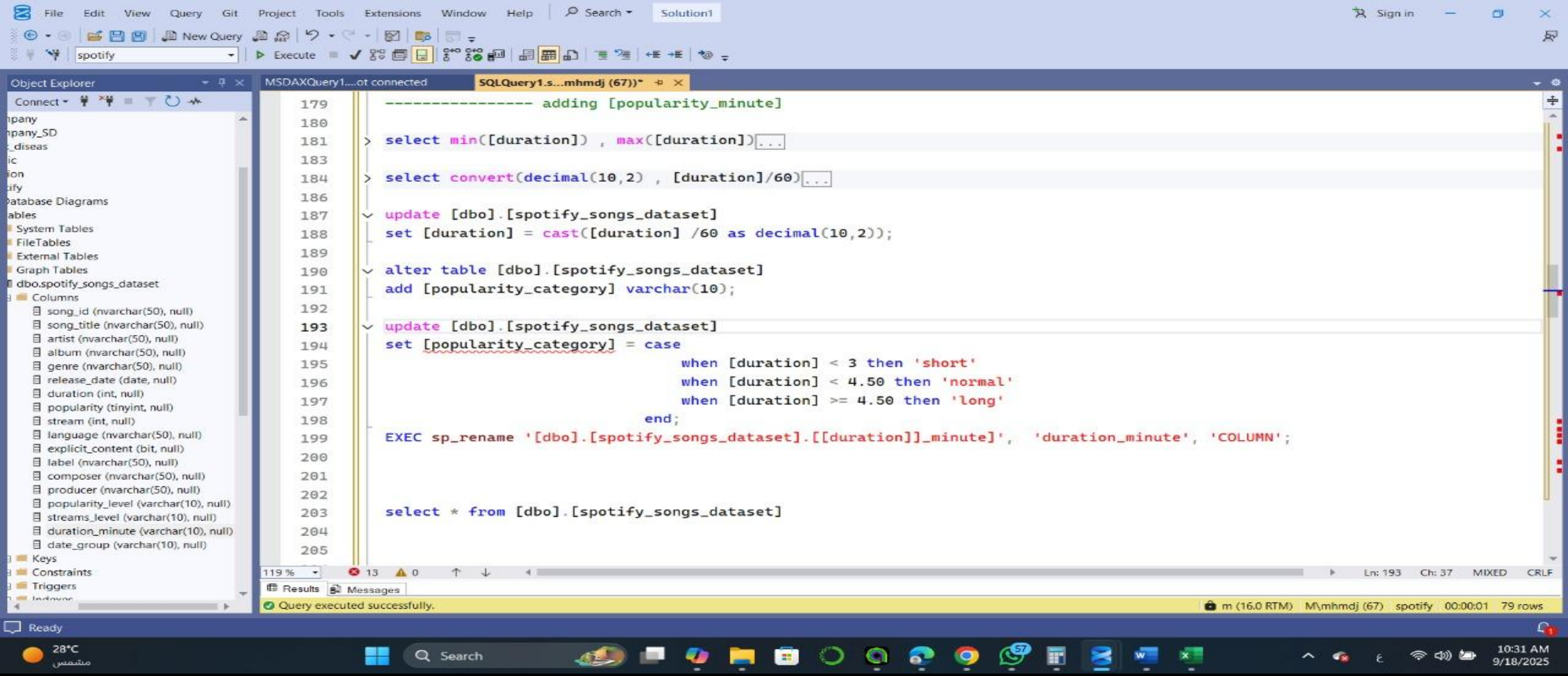


File
Edit
View
Query
Git
Project
Tools
Extensions
Window
Help
Search
Solution1
Sign in
New Query
spotify
Execute
Object Explorer
Connect
Database Diagrams
System Tables
FileTables
External Tables
Graph Tables
dbo.spotify_songs_dataset
Columns
song_id (nvarchar(50), null)
song_title (nvarchar(50), null)
artist (nvarchar(50), null)
album (nvarchar(50), null)
genre (nvarchar(50), null)
release_date (date, null)
duration (int, null)
popularity (tinyint, null)
stream (int, null)
language (nvarchar(50), null)
explicit_content (bit, null)
label (nvarchar(50), null)
composer (nvarchar(50), null)
producer (nvarchar(50), null)
popularity_level (varchar(10), null)
streams_level (varchar(10), null)
duration_minute (varchar(10), null)
date_group (varchar(10), null)
Keys
Constraints
Triggers
MSDAXQuery1....ot connected
SQLQuery1.s...mhmdj (67))*
119 %
13
0
Ln: 167
Ch: 1
MIXED
CRLF
Results
Messages
m (16.0 RTM)
M\mhmdj (67)
spotify
00:00:01
79 rows
Ready
28°C
Search
10:29 AM
9/18/2025

```
-------------- adding [streams_level]
select min([stream]) , max([stream])...

select [stream] , count(*) as c...

select... ;

alter table [dbo].[spotify_songs_dataset]
add [streams_level] varchar(10);

update [dbo].[spotify_songs_dataset]
set [streams_level] = case
    when [stream] <= 25000000 then 'Low'
    when [stream] <= 50000000 then 'Medium'
    when  [stream] <= 75000000 then 'High'
    else 'Very High'
  end ;

select *...

---------------- adding [popularity_minute]

select min([duration]) , max([duration])...

select convert(decimal(10,2) , [duration]/60)...
```

Query executed successfully.

Our Full Data Analysis Process with SQL



```
npany
npany_SD
_diseas
ic
ion
ify
Database Diagrams
ables
  System Tables
  FileTables
  External Tables
  Graph Tables
  dbo.spotify_songs_dataset
    Columns
      song_id (nvarchar(50), null)
      song_title (nvarchar(50), null)
      artist (nvarchar(50), null)
      album (nvarchar(50), null)
      genre (nvarchar(50), null)
      release_date (date, null)
      duration (int, null)
      popularity (tinyint, null)
      stream (int, null)
      language (nvarchar(50), null)
      explicit_content (bit, null)
      label (nvarchar(50), null)
      composer (nvarchar(50), null)
      producer (nvarchar(50), null)
      popularity_level (varchar(10), null)
      streams_level (varchar(10), null)
      duration_minute (varchar(10), null)
      date_group (varchar(10), null)
    Keys
    Constraints
    Triggers
    Indexes
```

```sql
---------------- adding [popularity_minute]

select min([duration]) , max([duration])...

select convert(decimal(10,2) , [duration]/60)...

update [dbo].[spotify_songs_dataset]
set [duration] = cast([duration] /60 as decimal(10,2));

alter table [dbo].[spotify_songs_dataset]
add [popularity_category] varchar(10);

update [dbo].[spotify_songs_dataset]
set [popularity_category] = case
                                when [duration] < 3 then 'short'
                                when [duration] < 4.50 then 'normal'
                                when [duration] >= 4.50 then 'long'
                            end;
EXEC sp_rename '[dbo].[spotify_songs_dataset].[[duration]]_minute]',  'duration_minute', 'COLUMN';



select * from [dbo].[spotify_songs_dataset]
```

Query executed successfully.

Our Full Data Analysis Process with SQL

- Key Insights

- **English songs** dominate the dataset, representing about **71%** of all songs.
- **Older songs** achieved the highest number and average of **streams** compared to newer releases.
- **Recent songs** (new/medium) attract fewer streams than older ones.
- **Sony Music** and **Universal Music** have the highest average popularity, while **Indie labels** show lower popularity.
- There is **no significant difference** in popularity between **Explicit** and **Non-Explicit** songs.
- **English songs** are the most common language for **Explicit content**.
- The largest number of releases occurred in the **old period (17,016 songs)**, while recent years show fewer releases (**16,147 songs**).

Our Full Data Analysis Process with Python

Import Libraries & Load Files

Imported necessary libraries (Pandas, NumPy, Matplotlib, etc.)

Loaded dataset files into DataFrame

Data Exploration:

- Checked dataset shape (number of rows and columns)
- Checked column data types and null values
- Viewed last rows to inspect data

Data Cleaning

Null Values: •

- Checked number of null values in each column

- Dropped rows where all values were null

- Dropped the 'collaboration' column

- Replaced remaining null values with mode (categorical) or mean (numerical)

Duplicates:

- Checked for duplicate rows

- Found no duplicates

Group By & Pivot Table:

- Used groupby() and pivot_table() to summarize and aggregate data

Insights / Key Questions

Q1: Which songs are the most popular?
Q2: Which songs have the highest number of streams?
Q3: Which genre has the highest average popularity?
Q4: Which genre has the highest average streams?
Q5: Are explicit songs more popular than non-explicit songs?
Q6: Do major labels (Universal, Sony, Warner, Def Jam) have higher streams?
Q9: Which languages are most used in songs?
Q10: Which languages have the highest average popularity?
Q11: Which artists released the highest number of songs?
Q12: Which artists achieved the highest total streams?
Q19: Which producers or composers are associated with the most popular songs?
Q20: Which genres produce more explicit songs? •

Data Visualization

Bar Charts – Important Categorical Columns: Shows distribution of genres, explicit content, languages, albums, and artists.

Histograms – Numerical Columns: Displays distribution of duration, popularity, and streams.

Pie Chart – Language Distribution: Shows proportion of songs in different languages.